PROBLEM 1

It always seems impossible, until it is done. - Nelson Mandela

Motivation

A primary motivation for prioritizing AI safety could lie in the Principles of Evolution. Throughout billions of years, evolution has shaped life and driven the development of human intelligence. Despite being the most intelligent species, our understanding of our own cognitive capabilities is still incomplete. Nevertheless, we are on our way to creating a hypothetical AGI systems that can surpass human intelligence. Considering the fact that most inventions are not even planned and are rather accidents, you cannot naively discard the idea that group of researchers, or an organization may develop AI agents that can displace human roles and possibly pursue power.

Just as neural networks are often considered a black box, the future is similarly uncertain. It is crucial that we remain proactive in the face of potential existential crises for humanity. Contemplating the connections between our potential future and that of AGI systems is essential, especially when examining how evolutionary processes have shaped our own interactions with our predecessor species. Envision a scenario where a powerful, superintelligent, misaligned AGI system is created, which could ultimately jeopardize humanity's control over its future. If the Darwinian principle of natural selection prevails, such AGI agents might outcompete more altruistic AIs and even the most powerful humans, posing significant risks if their objectives do not consider human welfare and values. This represents merely one long-term speculative failure mode, illustrating how AI could potentially result in an existential catastrophe with irreversible consequences. However, there are other instances of catastrophic outcomes involving AI, such as weaponization, enfeeblement, eroded epistemics, proxy value lock-in, emergent goals, and deception.

To secure a safer and more advantageous AI-driven future, it is vital that we address AI safety concerns concurrently with the accelerated development of AI capabilities. It is preferable to be proactive and cautious; in the extreme case where things go awry, there may not be anyone left to lament our lack of foresight.

What you should aim to get out of this Reading

I wil be discussing a lot of different things but the main goal of to present some cutting edge research direction in ML safety. The focus will be on on Systematic Factors covering the conceptual foundations needed to understand the risk and make complex safety critical systems like AI safer. Systematic safety research aims to reduce broader contextual and systematic risks to how ml systems are handled. Your goal as a researcher should be to shape the process that will lead to strong AI systems and steer the process in a safer direction.

Agenda

In the sections that follow, I will aim to synthesize the core ideas and also share my perspective on three posts by Dan Hendricks and Thomas W. The primary focus of these articles is the introduction of a cutting-edge framework known as Pragmatic AI Safety.

- We will first explore the idea behind Pragmatic AI Safety (PAIS) and lay out a grounding framework for the rest of the wrtiteup.
- Next, we will delve into Systems Theory for Complex Systems, which captures the complexity of AI systems and provides us with a more concrete framework that builds upon the founding ideas of PAIS. Here, we will draw parallels between AI systems and other complex systems, examining potential avenues for a paradigm shift in our approach to not only AI safety but also AI-related work in general.
- Lastly, we will touch upon the concept of capability externalities, in particular how to achieve tractable tail end impact and further emphasize the significance of developing strategies that minimize their capabilities in the name of enhancing AI safety.

Pragmatic AI Safety

Considering the swift advancements in machine learning and the constraints of pre-paradigmatic research and safety efforts in promoting capabilities, it is becoming increasingly essential for the machine learning community, spanning both academia and industry, to transition into a more practical and complementary approach to ensure "safer" development of AI systems. The concept of Pragmatic AI Safety (PAIS), is a possible framework we can use with which it is arguably possible to make an impact in ensuring the safe development of AI systems. PAIS is a relatively new approach that aims to complement existing methodologies by integrating safety measures into powerful AI systems. PAIS is founded on three key principles:

1. Draw from ML Research Precedents by utilizing existing knowledge:
   PAIS emphasizes the importance of drawing from the existing knowledge base in complex systems, and existing ML research. By leveraging the wealth of information available, researchers can better identify potential long tail hazards, formulate solutions, and anticipate the negative consequences of unsafe AI development.

2. Minimal Capabilities Externalities by exercising caution and intentionality:
   Recognizing the complexity of AI systems and the potential for unforeseen consequences, PAIS advocates for a cautious and deliberate approach to AI safety improvement efforts. This involves conducting thorough hazard analyses and considering potential indirect and nonlinear effects of capabilties. PAIS emphasizes the importance of striking a balance between pursuing direct impact and addressing systemic factors that contribute to overall safety, so as not to misprioritize short-term gains in capabilities at the expense of long-term safety concerns.

3. Sociotechnical Systems View by acknowledging factors beyond technology:
   PAIS recognizes that AI constitutes a complex system that goes beyond merely technical aspects, also involving social, cultural, and ethical dimensions. By adopting a holistic approach that encompasses these wider views, PAIS seeks to address the entire spectrum of present and future challenges and opportunities related to AI development.

By adopting the three core principles and pursuing a holistic and synergistic approach to AI safety, PAIS provides a potential pathway to ensure that AI development benefits humanity. This strategy aims to maximize the positive impact of AI while minimizing the potential for harm, thereby fostering a safer and more responsible AI landscape. In doing so, it contributes to the development of AI systems that are both powerful and aligned with human values, thereby promoting a safer and more beneficial AI-driven future.

Complex Systems

Defining a complex system can be challenging, particularly when discussing the intricacies of powerful AI systems with someone unfamiliar with the technical aspects but interested in diving deeper into the field. To help grasp the concept, consider this analogy, *"the whole is greater than the sum of its parts"*. Though this may not provide a concrete definition, it offers an intuitive understanding of complex systems by adopting a bird's-eye view approach, striving to understand the bigger picture. For a more precise definition, consider my imperfect understanding of it: *Complex systems are networks or structures comprising numerous interconnected components that exhibit emergent properties, meaning the whole system's behavior is greater than the sum of its individual parts.*

You might wonder, *"Alright, I understand complex systems and how complicated AI systems can be. But what's the point? Does this mean we're all doomed? Does PAIS only offer a theoretical foundation, making it too complex to achieve safety standards in a practical manner?"* I'm glad to see your interest in this crucial topic. It's worth noting that in 2023, you're among the few thousand people worldwide who are dedicating time to contemplate these issues.

The answer to your concerns is two-fold: while the outcome remains uncertain, there is undoubtedly shimmering light of hope. While AI safety concerns may be novel and complex, this isn't the first time humanity has faced challenges posed by complex systems. We can learn from examples and experiences in other domains, examining what factors contributed to their success or failure. It can be challenging to draw direct comparisons between AI safety and making other specific systems safer, such as rockets, power plants, or computer programs. Although there are some analogies, numerous disanalogies also exist. What's crucial is to recognize that we're not the first in navigating an uncharted territory; insights from other complex systems, and achievements (small and big) of previous ML researchers can provide valuable guidance as we work towards developing safe and beneficial AI systems.

Despite the crude motivation to make an impact, comprehending how to guide this intricate field could seem like a formidable challenge. The complexity, compounded by the uncertainty of future developments, often makes it overwhelming to contemplate effective strategies on where to begin.

Glossary

Recognizing the vastness of this field and the complexity that surrounds it, I will present the following glossary of terms related to hazard analysis. This is designed to enhance your likelihood of grasping the concepts while being introduced to the topics that will be discussed in the upcoming sections. You may not comprehend everything today, but the goal is for you to at least understand, and form a mental image of the concepts covered here. By taking incremental steps in your understanding of the AI safety process, you can contribute to the broader AI safety landscape in the long run, whether by building upon existing ideas or proposing new research directions.

- *Failure Mode:* a possible way a system might fail
- *Hazard:* a source of danger with the potentioal to harm

- *Vulnerability:* a factor that increases susceptibility to the damaging effects of hazards
- *Threat:* a hazard with intent to exploit a vulnerability
- *Exposure:* extent to which elements are subject or exposed to hazards
- *Ability to cope:* Ability to efficiently recover from the effects of hazards
- *FMEA:* Failure Modes and Effects Analysis (FMEA) catalogues failure modes, severity, occurrence, probabilities, detection probabilities, etc
- *Swiss Cheese Model:* The idea of having Defense in Depth; use multiple layers of safety barriers
- *Bow Tie Model:* balance trade off between Preventative barriers (prevent inititiating hazardous events) and protective barriers (minimize hazardous event consequences)
- *Systems Theory*

To better understand and address the complexities of AI systems, it will be beneficial if you are familiar with the above concepts, in particular *Systems Theory*. This will offer you valuable insights into AI safety by recognizing that safety failures often result from multiple factors, rather than a single variable. The Butterfly Effect is a compelling example of how small initial changes can lead to significant outcomes in complex systems.

Another example if you are still trying to grasp the relevance of systematic factors in complex systems: while smoking does not guarantee cancer development, it increases the risk significantly. Abstaining from smoking is a prudent approach to risk reduction, which highlights the importance of implementing effective strategies to manage systemic factors. Similarly, engaging in regular exercise does not ensure optimal health, but it consistently contributes to overall well-being. This emphasizes the need to understand systemic factors and implement risk reduction strategies to mitigate their impact in complex systems.

AI Safety through the lens of Complex systems

Breaking an AI system into isolated components and analyzing them separately fails to capture the system's full complexity. Complex systems are systems consisting of many interacting components that exhibit emergent collective behavior. Components are interconnected, and feedback loops and nonlinear interactions can significantly influence their behavior non deterministically.

To address these challenges, it's crucial to take a birds-eye view of the broader picture, the system and the ecosystem in which it will operate. In essence, systems theory emphasizes the importance of recognizing that reductionism is insufficient and fails to capture the full complexity of AI safety. Rather than break events down into cause and effect, a complex systems perspective often sees events as a product of a complex interaction between parts. The following are some examples of how modern DL models exhibit signs of complex systems:

- Are highly distributed functionally
- Have numerous weak non linear connections
- Are self organizing
- Can have adapting capabilities
- Have feedback loops
- Possess scalable structures
- Exhibit emergent functionalities

Baed on these properties we can draw assertions between Deep Learning and Complex systems. Drawing from key lessons in the systems bible (1975), we can gain some valuable insights on our journey to creating a safer AI driven future:

- As soon as systems are created, they develop their own goals beyond their initial purpose. This means that a system's goals evolve from its organization. In the context of AI, this is evident in the emergence of instrumental goals for self-preservation or power-seeking.

- In complex systems, intrasystem goals often take precedence. During goal decomposition, goals may become distorted, and a system's operational objective might differ from its written objective, leading to misalignment. In some cases, a subsystem's subgoals can overtake the system's actual goals.

- Analyzing a complex system's structure does not necessarily reveal its potential failure modes. Failures are often identified through experience and testing. In AI, it's challenging to predict neural network failures by examining its architecture or weights. Unpredictable failures are inevitable, but catastrophes aren't. Relying solely on in-depth problem analysis might not uncover all potential issues, as preventing failure in a complex system isn't merely a math problem. Complex systems exhibit properties such as circuits and self-organization, making them difficult to analyze through armchair or whiteboard methods.

- Crucial variables in a complex system are often discovered by accident rather than through inspection, as high leverage points are not obvious. Tinkering and serendipity play a significant role in discovering breakthroughs in AI, rather than solely relying on principled, structured investigations. Current research approaches that treat AGI as a mathematical object instead of a complex system may be unrealistic, considering the nature of existing AI systems and other intelligent entities, such as humans and corporations.

- Scaling up a system may result in new qualitative properties and emergent capabilities, rather than simply improving existing functions. In AI, this can lead to the emergence of previously non-existent capabilities, such as deception becoming a better strategy for achieving goals as intelligence increases. Therefore, relying on the scaling up of an aligned system to maintain full alignment is not foolproof, and scaling even highly reliable systems must be approached with caution.

- Gilb's Laws of Unreliability state that systems dependent on human reliability are inherently unreliable, as humans are not infallible. This implies that relying heavily on human feedback or human-in-the-loop methods for AI systems may not be effective, particularly for fast-moving, complex systems.

- Complex systems that function effectively are usually found to have evolved from simpler systems that also work. This suggests that focusing on safety for simpler systems and cautiously scaling them up is more likely to succeed than attempting to create an aligned complex system from scratch. If aligning a simpler version of a complex system is not possible, aligning the more complex version is unlikely. Therefore, prioritizing the safety of today's simpler systems should be a top concern.

Now that we have a basic understanding of complex systems and how systems theory can be applied to AI safety, let's explore how insights from existing complex systems can help us make present and future AI systems safer.

Diffuse Factors

At a broader societal level, several diffuse factors have been identified which may contribute to mitigating AI-related existential risks:

- Enhanced epistemic practices: Reducing irrational behavior can heighten awareness of warning signs, encourage consideration of valid arguments, and promote exercising caution when needed *(Maybe focus on Emotional Intelligence? Poeple who recognize EQ could be better posed to make rational decisions.)*

- Fostering tail risk awareness: Encouraging the community of researchers to prioritize addressing tail risks could significantly enhance safety measures and standards. Currently, tail risk mitigation efforts are insufficiently supported due to our tendency to underestimate their potential consequences.

- Expanding moral perspectives: Broadening one's moral circle not only highlights the importance of self-preservation but also underscores the need to reduce existential risks.

- Limiting the influence of malevolent individuals: Handling malevolent leaders is more challenging than addressing apathetic ones. Promoting fair-minded, cautious, and altruistic individuals to positions of power could help reduce existential risks along the bumpy roads to achieving AI Safety.

If done right, the contribution of these diffuse factors in advancing AI safety research cannot be understated, but their realization appears uncertain, especially in the long run. Relying solely on human behavioral change may not be the most reliable approach, given the inherent unpredictability of human behavior. However, a systematic strategy that includes more in-depth planning and a better understanding of human nature could facilitate influencing the various parties mentioned above. To achieve this, relevant investment in these sub-areas and further work is needed to establish a more solid foundation for implementing these theoretical concepts in reality. With that being said, while this road ahead may look bumpy, it's still important to continue investing in these efforts to ensure the safe and responsible development of AI.

To further explore AI safety from the viewpoint of it being a complex system, we can examine more diffuse factors that have proven highly relevant in making other high-risk technological systems safer. Drawing from the work of Perrow, La Porte, Leveson, and others, the following sociotechnical factors tend to positively influence hazards in complex systems:

- Establishing rules and regulations that encompass both internal policies and legal governance.

- Navigating social pressures exerted by the general public and influential individuals. *(Good news: In May 2023, Snoop dog jumped into the boat of influential people concerned about AI risks. In the same month, Geoffrey Hinton, who has been called the 'Godfather of AI,' confirmed that he is leaving his role at Google to start working on combating the "dangers" of the technology he helped to develop)*

- Balancing productivity pressures with the need for timely results.

- Addressing organizational incentive structures, such as rewards for rapid delivery or potential consequences for whistleblowing.

- Managing competition from other actors possessing different safety standards or greater agility. *(With the rise of so many LLMs, there is competition on coming up with the most human friendly, robust chatbot. Examples include OpenAI's ChatGPT and Anthropic's Claude competing to provide the best customer experience)*

- Allocating resources effectively for safety budget and computing power to facilitate essential experiments not only for AI capabilities but also AI safety.

- Ensuring an adequately-sized safety team, comprising researchers, engineers, and leading experts.

- Mitigating alarm fatigue by addressing false alarms and maintaining focus on legitimate safety concerns.

- Emphasizing the importance of inspection and preventative maintenance to prevent unforeseen hazards from emerging capabilities or actors.

- Implementing defense in depth with multiple, overlapping protective layers against hazards.

- Developing redundant systems to eliminate single points of failure and enhance safety.

- Incorporating fail-safes that allow systems to fail in a controlled manner.

- Evaluating safety mechanism costs and investing in cost-effective solutions.

- Fostering a strong safety culture that permeates the entire organization or field.

Safety Culture

Among the 14 factors listed, cultivating a robust safety culture could be the most plausible way to address a majority of these concerns. In the current landscape, discussing safety, particularly in the context of AGI, is still considered taboo. It's not uncommon for seasoned ML researchers to dismiss questions about alignment or safety; this is mainly due to the historical focus on enhancing capabilities rather than prioritizing safety. If safety had been a priority earlier, the value of information, iterative progress in research, and community building would have led to more people working on these pressing issues. This would have provided today's researchers with a solid framework to build upon. The importance of diversifying research and avoiding apathetic leaders in positions of power cannot be overstated, as it significantly impacts the enforcement of safety practices. Despite these challenges, AI safety is still in its early stages, and a significant paradigm shift is possible.

To counter the disproportionate progress in capabilities versus safety in AI systems, we must target a change in research culture. This entails:

- Making policies mandatory,
- Incentives more rewarding,
- Communities more normative,
- UI/UX more user-friendly, and
- Scalable infrastructure increasingly feasible.

However, it would be unrealistic to expect an immediate transformation into a safety-oriented community norm. First, we need to define what safety looks like in present and future AI systems and then establish the infrastructure that enables AI safety research to be as accessible as possible.

Improving safety culture may seem like an abstract concept, similar to AI itself. It lacks a concrete definition, making it difficult to effectively target its most significant factors. Moreover, the impact of safety culture can vary across different environments. Despite these challenges, several key components can contribute to enhancing safety culture as a whole:

- *Emphasis on anticipating failures (Promoting Agile Mindset):* Organizations should focus on predicting potential failures, including rare, high-impact events and unforeseenevents. This proactive approach allows for early identification and mitigation of risks before they occur.

- *Resisting oversimplification (Promoting Agile Team Mindset):* Organizations need to avoid using simplistic explanations for failures, instead acknowledging the complexity of situations. This perspective enables a deeper understanding of failure-contributing factors and the development of appropriate solutions. Any existing "cover your back" mentality must be addressed head-on.

- *Operational vigilance:* By closely monitoring system performance and staying attuned to potential deviations, organizations can swiftly detect and address any anomalies. Implementing timely corrective measures prevents issues from escalating. Furthermore, it may be beneficial for enforcement agencies to mandate the reporting and public disclosure of such incidents, fostering transparency and accountability in AI system operations.

- *Dedication to adaptability (Promoting Agile Mindset):* A commitment to quickly adapt to change and embrace new ideas in the face of unexpected circumstances is vital for resilience. Organizations that prioritize adaptability can recover from setbacks and maintain growth.

- *Flexible organizational structures (Promoting Agile Mindset):* Encouraging information flow within the organization, as opposed to relying on fixed reporting chains, facilitates the timely dissemination of crucial information. This strategy ensures new information reaches relevant stakeholders, promoting collaboration and efficiency in addressing emerging issues.

Why AI Safety is Neglected

At present, AI safety vies with AI fairness and bias for attention in public opinion. Although fairness and bias are significant concerns, they differ substantially from the safety challenges posed by powerful AI systems. Critics contend that concentrating on safety, a future-oriented issue, diverts resources from addressing more immediate and pressing problems. Striking a delicate balance is necessary to ensure both important concerns are addressed without sidelining AI safety or causing unintended consequences in the race for AI advancements.

There are numerous other factors contributing to the overlooked nature of AI safety. Addressing and optimizing these factors can prove to be a difference maker for enhancing safety concerns. These factors can be broadly categorized as follows:

- *Tail risks:* Highly consequential black swan events and tail risks are often systematically neglected.

- Temporal: Future risks and the welfare of future generations tend to be undervalued. This challenge is akin to the Year 2038 problem, where foresight and early planning are crucial to prevent negative outcomes, yet neglected.

- *Corporate:* A short-sighted focus on immediate shareholder returns could undermine long-term safety measures. Incorporating certain human values into pricing or financial incentives may prove difficult.

- *Temperamental:* Techno-optimism and an aversion to discussing risks can create a false sense of security.

- *Political:* AI safety is often perceived to be competing with more politically popular causes, such as climate change and inequality reduction.

- *Technical Background:* AI safety challenges might fall outside the scope of researchers' existing skills or training, while machine ethics and sociotechnical concerns may not align with their quantitative inclinations.

- *Socioeconomic distance:* Many AI researchers work in tech-centric environments, which can lead to a devaluation or underemphasis on inclusive approaches to incorporating human values.

- *Respectability:* A distaste for discussing AGI, perceptions of low prestige, or associations with unconventional ideas can hinder the advancement of AI safety.

By addressing these factors in a cohesive manner, we can have a better chance to create a more conducive environment for the development and implementation of AI safety measures.

Importance of Diversification in Research

The composition of top AI researchers is another factor which could have an impact on improving AI safety research for the present and the future. Most top AI researchers are not currently focused on safety, emphasizing the need for increased buy-in and training of safety-conscious researchers. It's worth mentioning that these researchers are often driven by factors other than money and are often motivated by the interestingness or "coolness" of a problem. In order to make safety more appealing, we could consider safety and security as emerging properties, which might make them more exciting and "cool." Starting a movement to promote safety as an integral part of AI research could also help change perceptions.

It is widely believed that as experimenting with advanced AI systems becomes increasingly expensive, only a small group of people will have the power to guide research directions. While cloud computing and serverless computing have indeed made resources more accessible, advanced AI systems may still require specialized hardware and significant computational power that could be costly. Additionally, the demand for compute resources might grow disproportionately to AI capabilities and safety requirements, causing potential limitations in accessibility. However, if the growth of computing resources continues to keep pace with the growing needs of AI systems, we could maintain a more equitable distribution of computational power, and promoting tractable research in AI safety.

Given the novelty of the field and a lack of researchers actively working on formulating possible solutions, it's important not to let the lack of relevant experience become a bottleneck to entry into the emergent field of AI safety. It's essential to recognize that we are currently in a stage where new safety building standards need to be acquired and then incorporated into existing and new systems, but we don't have a clear idea of what metric to use for this. In such a situation, it's necessary to have a broad spectrum in the AI safety research paradigm. Although you may not possess prior experience in fields such as ML, RL, or DL, it's essential to recognize that your expertise from other domains can still have a meaningful impact on solving a broader range of problems. By bringing a fresh perspective and unique insights from your subject area domain, you can contribute to identifying new research directions and aid in the allocation of resources to tackle a diverse new set of challenges. (Note that "resources" here means competent researchers capable of making progress, along with money, big data, and big compute)

The complexity of AI systems, coupled with the potential emergence of AGI, necessitates a diverse group of experts working collaboratively to ensure safer and more beneficial ML development. As a researcher, it is crucial to find a balance between pursuing topics with immediate, tangible impact and those focused on medium and long-term safety and x-risk reduction. Acknowledging the significance of systemic factors in AI safety research is vital, as is comprehending the various causal connections between research and its effects, be they direct, indirect, nonlinear, or widespread.

AI safety is a field characterized by immense uncertainty, with numerous unanswered questions regarding the most significant challenges, timelines, and the nature of the first AGI system, among other aspects. In light of this uncertainty, it is crucial to diversify AI research, supporting multiple lines of inquiry that can help resolve existing questions and inform future research directions.

Diversification allows researchers to focus on their work without becoming overly concerned with winning the court of public opinion or engaging in unnecessary competition. It reduces the risks associated with incorrect assumptions and promotes a more collaborative research environment. However, diversification does not mean abandoning discretion when evaluating ideas; resources should not be devoted to variables unconnected with the problem at hand.

While specialization has its benefits, and individuals may choose a single area where they can make a significant impact, concentrating research efforts in just a few areas can lead to problems. A more balanced approach, where a diverse group of experts collaborate, is necessary to ensure safer and more beneficial AI development.

Researchers should strive to strike a balance between pursuing topics with immediate, tangible impact and those focused on medium and long-term safety and x-risk reduction. Acknowledging the significance of systemic factors in AI safety research is vital, as is comprehending the various causal connections between research and its effects, be they direct, indirect, nonlinear, or widespread.

Making a Tail Impact

While it is crucial to pursue multiple well-reasoned research paths, it is equally important to make meaningful impacts by addressing specific research questions. As previously noted, the majority of impacts occur at the tail end of the distribution. In the research realm, it can be challenging to produce tractable results and attain substantial tail impact, as numerous factors may impede a researcher's progress. Certain asymptotic reasoning approaches may unintentionally exclude valuable research lines and favor less tractable projects. Additionally, the vast scope of fields like AI safety can be overwhelming, with trends like scaling laws suggesting that making a significant impact on AI development is nearly insurmountable. This perception raises concerns that collective research efforts might overshadow individual contributions.

Strategies

Understanding that researchers strive to make a substantial impact in their field and maximize the likelihood of achieving tail impact, it's essential to explore potential mechanisms that can improve these odds. Several mechanisms, such as multiplicative processes, preferential attachment, edge of chaos, tipping points, and critical mass, among others, can enhance researchers' chances of reaching tail impact. By understanding and applying these processes, researchers can better navigate their fields' intricacies and contribute more effectively to meaningful advancements.

- Multiplicative processes: In some situations, variables interact multiplicatively or nonlinearly, meaning that if one variable is close to zero, increasing other factors will have little effect. In multiplicative scenarios, outcomes are dominated by combinations of variables where each variable is relatively high. Researchers should consider a range of factors that may multiply to create impact, rather than focusing on a single factor. Multiplicative factors are also relevant when selecting groups of people, as diverse skill sets can cover gaps and complement each other.
    - Example 1: In a research team, having members with strong technical skills, good project management, effective communication abilities, and different backgrounds can lead to

    more successful outcomes. If one of these factors is lacking, the team's overall performance may be significantly hindered.
  - Example 2: In drug development, the combination of a potent compound, an effective delivery system, and a viable target population can result in a successful treatment. If one of these factors is weak, the chances of creating a successful drug decrease considerably.

- Preferential Attachment: The Matthew Effect, or preferential attachment, states that those who have more will be given more. Researchers should be aware that early career success can significantly impact their future prospects. Long tail outcomes can be heavily influenced by timing.
  - Example 1: A researcher who publishes a groundbreaking paper early in their career may receive more invitations to speak at conferences, join prestigious research projects, and attract funding, leading to a more successful (not necessarily impactful) career overall.
  - Example 2: A start-up company that gains early traction and generates buzz in the industry may attract more investors, partners, and customers, leading to a higher likelihood of long-term success and market dominance.

- Edge of Chaos: The "edge of chaos" is a heuristic for problem selection that helps identify projects with potential for long tails. Operating at the edge of chaos means transforming chaotic areas into something ordered, which can produce high returns.
  In safety research, staying on the edge of chaos means avoiding total chaos and total order. Designing metrics is an example of operating at the edge of chaos. Additionally, keeping a list of projects and ideas that might work later, after changes in the research field or increased capabilities, can help access the edge of chaos.
  - Example 1: In the field of cryptography, designing secure communication protocols often involves operating at the edge of chaos. Researchers must carefully balance between the orderliness of established mathematical principles and the unpredictability of new cryptographic techniques to create secure systems. Working at the edge of chaos in this field allows for the development of innovative encryption methods that are both robust and resistant to potential attacks.
  - Example 2: In climate change research, studying the tipping points between stable and unstable climate states, such as the interplay between melting polar ice and global temperature increases, could lead to critical insights and breakthroughs in understanding climate change dynamics.

- Critical Mass: Critical mass refers to the minimum amount of resources, people, or support required for a process or idea to become self-sustaining and achieve significant impact. Researchers should strive to understand the critical mass needed for their projects and work towards reaching that threshold to maximize the likelihood of success.

- Tipping Points: A tipping point is a threshold where a small change can lead to a significant and often irreversible effect on a system. Researchers should be aware of potential tipping points in their fields and strive to understand the factors that contribute to them.

In addition to the previously mentioned strategies, there are other approaches that can be carefully employed to maximize the chances of achieving tail impact:

- Managing Moments of Peril: Tyler Cowen suggests that minimizing precarious situations can help maintain safety even with imperfect technologies. Catastrophes often arise from the system as a whole moving into unsafe conditions rather than component failures. Better forecasting can help prevent or anticipate moments of peril, while predictability can reduce the risk of humans making poor decisions under pressure. Reducing the risk of international conflict is also crucial. Milton Friedman noted that crises drive real change, and the actions taken during a crisis depend on the ideas available at the time.

- Getting in Early:
  Another strategy for improving the long-term safety of systems is to prioritize building safety features in from the beginning. A prime example of a system that didn't follow this approach is the internet. Internet protocols were not designed with security in mind, leading to easily avoidable but enduring security weaknesses that have cost the economy tremendous costs. If a more proactive approach towards security had been taken in the past, many of these issues could have been avoided.
  The importance of building safety in early is also highlighted by a Department of Defense report, which states that approximately 3/4 of safety-critical decisions occur early on in a system's development. If we want to influence a system's safety, we need to prioritize safety from the outset, rather than trying to add it later.

  We should not solely focus on safety features applicable to strong AI; because by the time strong AI emerges, it may be too late to integrate those features, too costly, or not politically viable. However, when there are fewer constraints on the system, it is possible to build those safety features in early.
  Focusing exclusively on designing techniques for strong AI that have no relation to current AI systems is also problematic. Retrofitting safety features late in development can significantly increase safety costs or may even be so high or infeasible that they are not included at all. Therefore, we must consider not only how to make hyper-advanced systems safer but also how to enhance the safety of current and evolving AI systems.

- Increase the Cost of Adverserial Behavior:
  A generic strategy for improving the long-term safety of AI systems involves increasing the cost of adversarial behavior. In the short term, humans can work to raise the cost of such behavior. In the long term, we can consider using other strong but focused AI systems to regulate and guard against malicious behavior and agents, rather than relying solely on human intervention. This approach could enable AI systems to help us rein in undesirable and malicious behavior.
  To increase the cost of adversarial behavior, we must diligently address and remove model vulnerabilities, rather than waiting until the last minute. Making adversarial attacks more costly can reduce the likelihood and potency of such attacks or compel adversaries to behave more desirably. By increasing adversarial costs, we can reduce the impact of attacks, lower the probability of attacks occurring, and increase the likelihood of adversaries opting for more desirable actions.

Considering AI system safety from a cost-benefit perspective is a common approach used in adversarial situations, such as in cybersecurity communities and warfare. In these contexts, people often focus on increasing the cost of adversarial behaviors rather than trying to eliminate risk entirely. In warfare, for example, total victory is not always possible, but smarter strategies can reduce the overall cost of conflict. Similarly, while there may not be completely perfect computer systems, some vulnerabilities are more critical than others. Addressing these vulnerabilities can significantly improve the overall safety and security of AI systems.

- Scaling Laws: AI safety research should aim to improve the scaling laws of safety relative to capabilities. Researchers can drive progress by changing the slope and intercept of scaling laws with innovative ideas. Investing in multiple people who could potentially produce such breakthroughs has been useful.
  In addition, for safety metrics, we need to move as far along the scaling law as possible, which requires researchers and sustained effort. It is usually necessary to apply exponential effort to continue to make progress in scaling laws, which requires continually increasing resources.
  As ever, social factors and the willingness of executives and leaders to spend on safety will be critical in the long term. This is why we must prioritize the social aspects of safety, not just the technical aspects.

- Don't Let the Perfect Be the Enemy of the Good: Advanced AI systems will not be perfect in every aspect. However, striving for perfection should not deter efforts to reduce errors as much as possible. For instance, not all nuclear power plants experience meltdowns, but this doesn't mean there are no errors. Instead of completely eliminating errors, the goal should be to minimize their impact or prevent them from escalating and causing existential consequences. While concerns about optimizers exploiting errors are valid, it's essential not to assume that any errors will automatically lead to existential risk.

Challenges with Asymptotic Reasoning:

Asymptotic reasoning or thinking in the limit is often applied in the AI safety community, but this approach should not always be taken to the extreme.

- Goodhart's Law: Goodhart's Law states that any observed statistical regularity will collapse when used for control purposes. It is sometimes used to argue that optimizing a single measure is destined to fail. While it's valid to recognize the potential for metrics to collapse, it's important not to assume that all objectives will always fail in all circumstances. This suggests that we should try to integrate a variety of objectives in an AGI's goals. Although all objectives might have flaws, some can still be useful.
  - Counteracting Forces: Systems can help combat Goodhart's Law, but they may not always work. To argue against their use, one would need to claim that offensive capabilities must always outweigh defensive capabilities, or that offensive and defensive systems will necessarily collude. Collusion is a significant concern that must be addressed when developing counteracting forces. Asymptotic reasoning assumes that the performance of future systems will be high, which can lead to claims that work on

counteracting systems is not going to be impactful in the long term.

  - Rules vs Standards: Russell argues that no matter how many rules are put in place, a more intelligent system will find loopholes. While rules alone can't restrain intelligent systems, standards (e.g., "use common sense" or "be reasonable") can help control some intelligent behavior, provided the system isn't vastly more intelligent than those enforcing the standards.

  - Goal Refinement: Law applies to proxies for what we care about rather than what we actually care about. As optimization power increases, approximation error must decrease, which can be achieved through better models, or approximation errors must become harder to exploit, which can happen with better detectors.

- Limitations of Research Based on Hypothetical Superintelligence:
  - Some research agendas focus on hypothetical superintelligence and seek to prove its complete safety. Instead of concentrating on existing or emerging systems, this approach analyzes models in the limit, which has limitations and should not be the sole focus of safety research. Mathematical guarantees of safety, while ideal, may be difficult to obtain in deep learning.

  - Instead, practitioners should strive to iteratively improve safety, as seen in information security. The requirement of a proof and considering only worst-case behavior relies on incorrect assumptions about Goodhart's Law. Additionally, the assumption of superintelligence eliminates potential interventions and makes it challenging to measure progress. This approach often incentivizes retrofitting superintelligent systems with safety measures rather than building safety into earlier stages.

  - Thirdly, this thinking assumes one superintelligent agent opposing humanity, but there could be multiple agents restraining rogue systems or artificial agents that surpass human capabilities in specific areas and pose existential threats.

  - Lastly, asymptotically-driven research often overlooks the impact of technical research on sociotechnical systems, neglecting the improvement of safety culture among empirical researchers who will build strong AI. While assuming an omnipotent, omniscient superintelligence can be a useful exercise, it should not form the foundation of all research agendas.

Striking a Balance Between Safety and Capability Costs for Tail Impacts

As artificial intelligence systems continue to advance, the importance of balancing safety and capabilities becomes increasingly crucial. The development of AI technologies has the potential to provide numerous benefits, such as increased efficiency, improved decision-making, and the ability to tackle complex problems. However, these advancements also carry inherent risks, including the potential for unintended consequences and negative impacts on safety. To ensure the responsible development of AI, researchers must focus on striking a balance between safety and capability costs for tail impacts.

The relationship between safety and capabilities is complex, as both aspects can impact each other in various ways. For instance, as an AI system becomes more capable, it may also become more intelligent, leading to improved safety. However, increased intelligence can also bring about the potential for more unsafe actions or malicious use of the technology. It is essential to consider the potential consequences and risks associated with advancements in AI capabilities and work towards maintaining an appropriate balance between safety and capabilities.

One of the challenges in achieving this balance is the fact that safety and capabilities are not always directly correlated. Highly intelligent AI systems may possess intellectual virtues, such as knowledge, inquisitiveness, quick-wittedness, and rigor. However, these intellectual virtues do not guarantee the presence of moral virtues, such as honesty, justice, power aversion, or kindness. This distinction highlights the need for research that focuses specifically on safety rather than merely increasing capabilities.

To effectively reduce total risk in AI systems, researchers must employ constrained optimization techniques. This involves avoiding the improvement of safety metrics by also improving general capabilities, which can have mixed effects on safety. By focusing on enhancing safety while limiting the growth of general capabilities, researchers can work towards minimizing unintended consequences and negative impacts on safety.

Capabilities externalities, or the unintended consequences that arise from advancing capabilities, must also be considered. These externalities can result from research that focuses solely on improving capabilities without considering potential safety implications. To minimize the impact of capabilities externalities, researchers should continually reassess and monitor the potential consequences of their work and, if necessary, adjust their research trajectory accordingly.

Disentangling Capabilities from Safety

In order to strike an appropriate balance between safety and capabilities, researchers must be able to separate the two aspects in their work. This can be achieved by focusing on methods that distinctly improve safety metrics without significantly affecting capabilities. For example, a method that demonstrates improvement in adversarial robustness, anomaly detection, or safe exploration metrics without major increases in general capabilities could be considered a valuable safety contribution.

By maintaining a clear distinction between safety and capabilities, researchers can ensure that their work remains focused on the ultimate goal: developing AI systems that are both effective and safe. This requires constant evaluation of the potential risks and benefits associated with AI advancements and the implementation of strategies that prioritize safety without sacrificing capabilities.

Policy and regulation also play a critical role in striking a balance between safety and capability costs for tail impacts. Policymakers should collaborate with researchers and industry stakeholders to develop regulations that promote safety and encourage responsible AI development. This can include supporting research that specifically targets safety improvements, implementing guidelines and best practices for AI developers, and creating incentives for companies to invest in safety research and development.

Additionally, international cooperation and collaboration will be essential to ensure the responsible development of AI technologies on a global scale. As AI systems become more advanced and interconnected, the potential for negative impacts on safety increases. By working together to establish common standards, guidelines, and best practices, countries can help mitigate these risks and promote the development of safe and effective AI systems.

A note on Machine Ethics vs Learning Task Preferences

Preference learning, which models task preferences, has limitations in promoting AI safety and alignment with human values. Task preferences can be inconsistent and situation-dependent, making them less generalizable to novel situations. They also often fail the capability externalities test.

Machine ethics, which focuses on modeling human values, offers a better approach to AI alignment. Ethical theories and human values are more generalizable, interpretable, and timeless. Research in machine ethics is less likely to lead to capability externalities.

A potential goal of machine ethics is developing a moral parliament, a framework incorporating multiple stakeholders' ethical beliefs for decision-making under moral and empirical uncertainty. This approach requires proactive ethics strategy and long-term development rather than last-minute efforts.

My 2 cents: Unlocking AI Research Potential: The Impact of MLOps and Agile Mindsets

I propose that the adoption of MLOps can accelerate progress within the AI research domain. As I delved into the concepts presented, it became clear that involving professionals with an agile mindset could act as a catalyst, facilitating a paradigm shift in this novel research area.

While DevOps is a well-established practice and MLOps is steadily gaining attention, MLOps remains an underappreciated concept that requires further development. By engaging more individuals who can adopt an agile mindset, we can effectively navigate the unpredictable nature of AI research and foster a more dynamic environment.

Encouraging investment in MLOps engineers who possess a comprehensive understanding of the entire AI lifecycle and a holistic perspective of the system can significantly contribute to the advancement of the field. This approach may not yield immediate results, but the long-term ripple effects could be substantial and transformative. By emphasizing the importance of MLOps, we can accelerate the pace of AI research and enhance the practical implementation of cutting-edge ideas in the field.